神经网络BP算法与回归分析算法进行统计预测的比较研究

庞晶　乔洪宾

内容提要　人工神经网络是80年代迅速兴起的一门科学,目前,对其各种基本算法研究很多,但对算法的实际应用却研究的较少。本文试运用人工神经网络中关于多阶层神经网络及误差逆传播算法(Back－Propagation),简称BP法的相关知识,对因果关系类的统计数据进行预测,并与传统回归分析预测方法的结果进行比较,试图将神经网络的算法应用于统计预测中。

关 键 词　人工神经网络(ANN)　BP算法　连接权

一、引　　言

人工神经网络是人工智能方面迅速兴起的一门科学,它是模拟人脑结构和激励行为的并行非线性系统,它由大量的称为神经元的简要信息处理单元构成,大量性能简单的神经元可组成一个结构复杂、性能完善的系统以完成各类复杂任务,整个网络的信息处理是通过这些神经元的相互作用完成的。

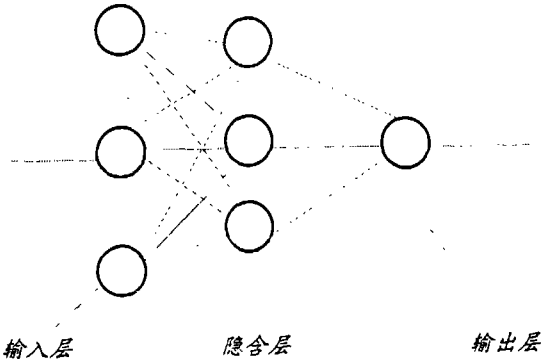
考虑到神经网络具有按相似输入产生相似输出的联想推理功能,且对因素的相关性要求较低,本文将利用某系统的前几期有关数据进行网络的学习,利用神经网络的非线性映射与联想推理能力,进行预测。成功的网络模型,就可以作为类似的统计问题的预测模型。

二、误差逆传播算法(Back－Propagation)

(一)BP算法

误差逆传播算法(Back－Propagation)是目前应用最为广泛的网络之一,其核心就是把一组样本的输入问题变为一个非线性优化问题,使用了优化中最普通的梯度下降法,运用迭代运算求解权,加入隐结点使问题的可调参数增加,从而得到更精确的解。

典型BP网络是三层前馈网络,即输入层、隐含层和输出层。本文采用的网络图如下:



当一对学习模式提供给网络后，神经元的激活值从输入层、隐含层向输出层传播，在输出层各神经元获得网络的输入响应之后，按减少希望输出与实际输出误差的方向，从输入层、经各中间层逐层修正各连接权，最后回到输入层，故得名“误差逆传播算法”。

网络中的响应函数采取 S 形函数(Sigmoid)：

$$f(x) = 1/[1 + \exp(-x + \theta)]$$

其中 θ 为阀值。因为 S 形函数更接近于生物神经元的信号输出形式，所以选用 S 形函数作为 BP 网络的输出函数。

(二)BP 网络结构和参数的选择

BP 网络的输入层与输出层单元数是完全根据问题的需要来设计的，对于隐含层单元的选择，目前还没有现成的式子可用。如本文是选用输入层为 5 个单元，输出层定为一个单元。而中间层的维数则要根据其它参数选定后经多次反复计算才能确定，本问题最终定为 7。

一般来说，任何一个三层的网络，即只含有一个隐含层的 BP 网络均可以逼近一个闭区间上的连续函数，也就是说，任何一个三层的网络可以完成任意的 N 维到 M 维的映射，因此本文就选用只含有一个隐含层的 BP 网络。

初始权值对于学习能否顺利关系较大，由于系统是非线性的，一个主要的原则就是希望初始权在输入时使每个神经元的状态值接近于 0。这样可保证网络一开始不至于落到那些平坦的区域上，导致网络局部最小。因为 BP 算法是采用了按函数梯度下降的方向进行收敛的，初始权值的选择不当，很可能使网络学习失败。

三、统计预测实例

(一)问题的提出

统计预测模型可分为因果关系模型和时间序列模型，一般回归分析是进行因果关系分析及预测的有效手段。因果关系所研究的问题实际上是总量指标与有关因素指标在变化过程中的数量关系。

以某地区城乡居民收入差预测为例，数据见下表

表1　　1980－1989年城乡居民收入差距及影响因素

年份	城乡居民收入差	二元结构系数	农产品价格剪刀差	城市居民隐性收入比重	农民非农产业收入比重	城乡人口比例
1980	3.09	9.05	0.29	0.226	0.088	0.241
1981	3.02	8.08	0.267	0.206	0.107	0.253
1982	2.74	7.45	0.244	0.203	0.139	0.268
1983	2.44	7.28	0.235	0.207	0.165	0.276
1984	2.39	7.02	0.210	0.194	0.182	0.300
1985	2.26	7.06	0.205	0.184	0.216	0.311
1986	2.60	6.98	0.18	0.163	0.207	0.325
1987	2.64	6.92	0.167	0.165	0.216	0.339
1988	2.49	6.88	0.185	0.165	0.272	0.348
1989	2.73	7.32	0.212	0.163	0.280	0.355

注:资料来源《统计研究》1995年第6期第52页。

为使本文研究方向更明确、更具有可比性,使问题简化。我们选用了《统计研究》1995年第6期中马恒运同志的一篇文章,题为《如何进一步分析逐步回归分析剔除的自变量》,该文是用回归分析的预测方法,建立模型,预测了某地区城乡居民收入差的趋势。本文将使用神经网络中的BP算法,也对表1中的数据进行处理,然后与原文中现成的数据和结论进行比较,试图从中找出快速提高BP算法精确度的方法。

一般的作法是应用数学中的统计知识,对以上数据作非线性回归,作出拟合曲线来预测。显然回归模型是进行因果分析及预测的有效手段。在建立模型时首先将上述各变量引入模型进行逐步回归和参数分析显著性检验,通过比较和筛选后,才能获得模型,而且在拟合过程中,还要对自变量进行相关性分析计算。如"农产品价格剪刀差"一项,就被作为不显著的因素剔除在模型外,但从理论上讲似乎是不可能的。因为我国经济的发展是以维持一定的农产品价格剪刀差为代价的,它肯定是造成城乡居民收入差距主要因素之一,这样必然影响模型解释和预测功效。

如何来解决这种数学方法和经济理论上的矛盾,以及如何充分地使这些不显著自变量与显著自变量一起参与模型的经济分析过程,的确是个难题。而神经网络BP算法的特点恰是解决这个难题的优势,因为神经网络在学习过程中自动地将各因素之间的相关程度以权值的大小体现出来,无须进行参数分析、显著性检验等繁杂手续,直接切入主题,因而能较好地解决这一矛盾。

(二)网络模型及参数的构造

示例的目的是根据以往有关数据来预测以后几年的城乡居民收入差。本文把表1中每一行,既每年的各种因素的结果看作一个样本(或模式),并把第二列的收入差作为每个样本的期望输出值。将上述表中数据输入计算机进行计算。

下面分步构造出这个问题对应的网络模型。

根据问题的性质可定出输入层数为5个单元,即表中第三至第七列作为网络的输入值;输出层单元数为1,表中的第二列(收入差)作为目标输出值;由网络的原理可知隐含层定为1即可,隐含层单元数须根据具体情况而定,由经验公式知其下界为3,通过几次运算及考虑到网络的稳定,所以先取为5,最后经调整定为7,网络的结构为(5-7-1)。

表1中收集的原始数据不能直接用来训练网络,因网络的输入项要求在-1到+1之间,因此对于本模型所有数据均除以所有数据中最大值M与最小值m之和,对于任意一个输入值S应该变为

$$S(j) = S(j)/(M+m),$$

这个过程本身就是对样本集的规律作出一些优化,去掉相同的部分,更有利网络的运行,为便于运行,已将该部分编入程序。

BP方法的实质是通过对由下而上的训练集中特殊模式进行学习来掌握普通的规律,最后达到认识大自然无限多的模式。所以要求这些模式具有比较典型的能说明这问题共性的特征。

在上述过程中,可以看到随着学习时间的增加,学习效果有所提高,可达到顶端水平,即误差达到极小,此时网络开始到过渡训练,测试效果有所下降,在这个最优过程中会出现几个自由度,包括学习率、动态参数、隐单元数、层数和训练集的构成。

BP算法的步骤如下:

Step 1 初始化　给各连结权及阀值赋予(-1,+1)间的随机数。

Step 2 随机选取一模式提供给网络,并输入期望输出。

对于每次训练输入应是新的值,直至权稳定为止。

Step 3 计算实际输出值,满足误差要求即可结束运算,否则继续。

Step 4 改写、调节权值。

Step 5 返回Step 2。

(三)网络参数的选定及数据预测

该网络的模式之间差别不大,它要求系统学习率的初始值选得不易太大。如本例开始取0.5还是可行的,随着学习过程的进行,逐渐减小到0.25.见表2。

表2　　BP算法模型最终计算的有关参数值

结构	输入层	5	参数	学习率	0.25
	中间层	7		记忆率	0.75
	输出层	1		训练次数	6000
规　模	样本数	10		全局误差	0.002

用训练好的模型分别对1990—1993年某市的收入差进行预测,得出结果(见表3)。

表3 BP算法结果相对误差表

年份	实际收入差	BP算法	
		收入差	相对误差%
1990	2.84	3.04	7.14
1991	2.92	2.85	2.40
1992	3.05	3.09	1.31
1993	3.07	2.92	4.89

从总体效果来看,BP算法的预测结果还是比较令人满意的,相对误差也不大,但比回归模型的预测结果特别是在稳定性上要差一些,我们分析主要原因是因为网络的训练和测试的不够好,对于参数选取及样本的特征抽取不理想。我们认为它毕竟是一种新的预测方法,存在着一些不足也是难免的,我们相信,随着经验的积累对网络模式的调整将会越来越好。

四、神经网络与统计方法的比较

许多因果关系问题的预测模型十分复杂,还要对自变量先行预测。而用神经网络的BP算法来预测,则会使问题简化,直接切入主题,更具有实际使用价值。

神经网络较之一般的统计方法约有以下优点:

在统计方法里,我们通常必需了解哪些因素对所要解决的问题是相关的,而神经网络则不需要考虑着一点,因为在神经网络的学习过程中自动地将相关程度以权值的大小体现出来。统计方法是以一种间接的方法来体现相关性的,必须经过五步迂回过程,而神经网络则要直接得多。神经网络模型能同时处理几百种因素,其中某些因素与问题求解只有很小的相关性,但将它们集合起来考虑,则可使得困难问题的求解更加精确,这是传统的统计方法所不能及的。

当然误差逆传播神经网络并不是一个十分完善的网络,它存在以下一些缺陷:

学习速度太慢,即使一个比较简单的问题,也需要几百次甚至上千次的学习才能收敛;不能保证收敛到全局最小点以及存在所谓"局部最小值";网络隐含层的数目与隐含层单元的选择尚无理论上的指导,而是根据经验确定。

本文通过实例分析,对BP法模型中规模的确定,学习率及动态参数的选择和模式集的分布作了些有益的尝试。我们认为该方法能处理高度非线性问题,具有自学习、自组织能力,具有良好的稳定性,它是神经网络的又一个较成功应用的实例。当然,本文的研究工作还很肤浅,还有许多方面的工作有待继续完成,如在权数的分配上还没有解决清楚,在模型的预测误差方面还须进一步完善。

(作者单位:内蒙古工业大学基础部、管理工程系)

参考文献

①焦李成:《神经网络系统理论》,西安电子科技大学出版社,1990 年版。
②史忠植:《神经计算》,电子工业出版社,1993 年版。
③Jones, W. W., Backpropagation, BYTE, Oct, 1987。
④R. Hecht - Nielsen, "Theory of Backpropagation Networks" Proc. IJCNN - 89, I - 593, 1989。
⑤王伟:《人工神经网络原理》,北京航空航天大学出版社,1995 年版。
⑥王文剑:《用神经网络方法进行暴雨预测》,参见《第三届中国人工智能联合学术会议论文集》,1995 年版。
⑦马恒运:《如何进一步分析逐步回归分析剔除的自变量》,参见《统计研究》,1995 年第 6 期。
⑧华伯泉:《经济预测的统计方法》,中国统计出版社,1988 年版。
⑨Borland International, Inc, BORLAND C + +, Version 3.1, 1991。